INVITED PAPER

Vision of a Visipedia

This paper envisions bringing together the knowhow and hard work of computer vision researchers into an online tool to form a repository of human understanding of visual imagery.

By Pietro Perona

ABSTRACT | The web is not perfect: while text is easily searched and organized, pictures (the vast majority of the bits that one can find online) are not. In order to see how one could improve the web and make pictures first-class citizens of the web, I explore the idea of *Visipedia*, a visual interface for Wikipedia that is able to answer visual queries and enables experts to contribute and organize visual knowledge. Five distinct groups of humans would interact through Visipedia: users, experts, editors, visual workers, and machine vision scientists. The latter would gradually build automata able to interpret images. I explore some of the technical challenges involved in making Visipedia happen. I argue that Visipedia will likely grow organically, combining state-of-the-art machine vision with human labor.

KEYWORDS | Crowdsourcing; image understanding; machine learning; machine vision; Visipedia; visual recognition; Wikipedia

I. DIGITAL DARK MATTER

The world wide web is a recent invention; however, the need for something like it has been felt for a long time. In a 1945 article titled "As We May Think" and published in *The Atlantic Monthly* Vannevar Bush lamented [6]:

> The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships.

Bush had a solution in mind: the *memex*, a "device for individual use" in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with speed and flexibility. After exploring a number of technical ideas that would make his memex feasible, he reflected:

> All this is conventional, except for the projection forward of present-day mechanisms and gadgetry. It affords an immediate step, however, to associative indexing, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the memex. The process of tying two items together is the important thing.

Bush's memex is widely regarded as foreshadowing hypertext and the world wide web. The web is, of course, not only "a device for individual use"; it is shared by all humans, and this makes it vastly more useful. Surely Bush's vision has been thus exceeded by a large measure.

Or, has it? Bush observes that "much needs to happen" between data collection and final use of the data. This is why search engines, which index web content automatically and make it available to us in a fraction of a second, are as important as the information that is available. Indeed, whatever is not properly cross referenced, indexed, and hyperlinked ends up lying fallow as "digital dark matter" on our hard drives. It is there, but we cannot easily access it. This is currently happening to the largest segment of the data we collect and store: pictures. We are now gathering and storing mind-boggling numbers of digitized photographs, drawings, diagrams, videos, and movies of all sorts, in astronomy, biology, medicine, physics, and engineering. Beyond science and engineering, even more photographs, videos, and digitized films are being collected by individuals and organizations for entertainment and commerce. However, as I will discuss in Section II, pictures are dark matter: by and large, they cannot be searched, they are not hyperlinked, and they are a giant digital missed opportunity.

Manuscript received May 4, 2009; revised September 17, 2009; accepted April 9, 2010. Date of publication June 3, 2010; date of current version July 21, 2010. This work was supported by the California Institute of Technology (Caltech) and by the Office of Naval Research (ONR) University Research Initiative (MURI) under Grant N00014-06-1-0734.

The author is with the California Institute of Technology, Pasadena, CA 91125 USA (e-mail: perona@caltech.edu).

Digital Object Identifier: 10.1109/JPROC.2010.2049621

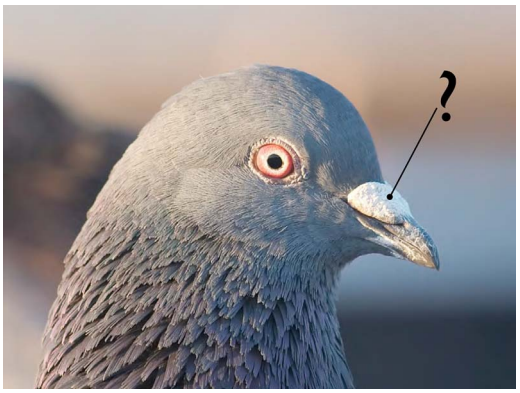


Fig. 1. *The mystery fleshy bit at the base of a pigeon's beak. What is its name? Why is it there? (Adapted from "Columba Livia," photograph taken by Dori. Available on Wikimedia as Pigeon_portrait_4861.jpg.)*

I would like to explore here what could be done to make pictures first-class citizens of the web. I will discuss a new concept, the "Visipedia," with the purpose of highlighting what could be different: machines could be able to interpret images and videos as much as we do, they could ask and answer "visual" questions, and they could collaborate with humans at gathering and organizing visual knowledge, connecting seamlessly text to images, images to text, and images to images.

II. THE JOYS AND SORROWS OF SEARCHING WIKIPEDIA

Wikipedia is one of the great surprises and success stories of the web. Both the quality and the quantity of the information that may be found in Wikipedia have grown by leaps and bounds since its founding in 2001 (see [878d8][1]). Anyone with an internet connection can access the most up-to-date version for free. Many of us have come to rely on Wikipedia as the first and primary source of information on almost any topic. Some of us use it routinely in teaching, alongside textbooks and class notes. A few of us have contributed our expertise to Wikipedia, possibly affecting more people than we do by teaching in the classroom. What a wonderful creation it is!

And yet, there are frustrating moments. I was recently sipping cappuccino in *Piazza delle Frutta* in Padova, Italy, and observing the pigeons peck at the crumbs that I had let fall from my brioche. I was suddenly made curious by the fleshy bit that one can see at the base of a pigeon's beak (see Fig. 1). Why is it there? What is its name? If a friend, knowledgeable about birds, had been sitting next to me at that time I would have pointed my finger and asked "What is that thing?" Alas, no such friend was at hand. I made a mental note that, once at home, I should consult Wikipedia and find out.

Here is how things went at home. I typed "pigeon" into Wikipedia's search box and I was redirected to the "Columbidae" page [2opkha]. It is a long page, the fleshy bit is visible in many of the pictures. Of course, I could not click on those and ask "what is this?" Fortunately, the page has a "morphology" section, but I could not find the information I was looking for. So, having given up on "Columbidae," I moved to the "beak" page [5b6xs] which contains a beautiful schematic of different types of beak (see Fig. 2). Unfortunately, none of those shows the pigeon's fleshy bit, and they are not clickable anyway. So, I decided to trawl through the text of the "anatomy" section of the page. One sentence attracted my attention:

> The nares are usually located directly above the beak. In some birds, they are located in a fleshy, often waxy structure at the base of the beak called the cere (from Latin cera).

Might "cere" be the thing? I clicked on the word and, bingo!, the prominent picture of a pigeon's head, fleshy bit included, told me that I had hit the jackpot. I went back to the "Columbidae" page and searched for "cere" and, there it was, towards the beginning, but I had missed it in my first scan through the page.

Was looking for "cere" a pathological special case? No. There is worse. My father-in-law recently saw a mushroom during a stroll. He snapped a picture and e-mailed it to me asking for help in identifying it (Fig. 3). I happened to know the mushroom, *Amanita Pantherina*, and it has a page in Wikipedia, but it would have been virtually impossible to find it in Wikipedia starting from the picture as the only clue. This is exactly why my father-in-law sent me a photograph. Wikipedia's "mushroom" page, by the way, recommends that if you wish to identify a mushroom you should use a reputable printed field guide.

What do we learn from these examples? First, often the relevant articles *are* there. They are just difficult to find if you do not have a keyword to start with. If I had known the word "cere" and had wanted to know what it is, it would have been a matter of seconds to find out: both a verbal description and pictures of it. The converse, having a picture and looking for the information that goes with it, is instead difficult. Is there something intrinsically difficult with using pictures as query keys? No. I can point at a picture and any knowledgeable human can give me the information instantly.

There is another limitation in Wikipedia, and it is even more serious than not being able to carry out visual queries. Look, for example, at the page on "liver" [oaof4]. The text is quite informative. On the right you will see a (nonclickable) image of a sheep's liver, illustrating its anatomical landmarks; below it is a drawing of a human abdomen, showing the main organs and the liver; further below there is

[1]This is a tiny URL, thus [878d8] means http://tinyurl.com/878d8, which expands to http://en.wikipedia.org/wiki/Wikipedia.

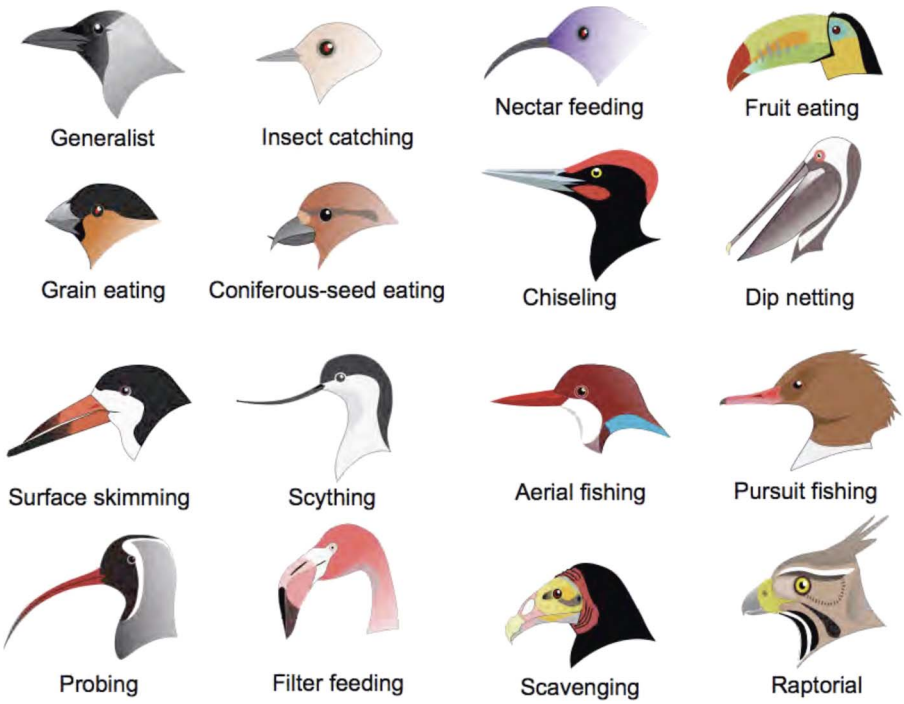


Fig. 2. ***Wikipedia page on "beak" contains this lovely and informative picture. Unfortunately, the picture is not clickable: it is a dead-end. Why has nobody taken the time to make it clickable? Where could it lead us if it were clickable? (Adapted from L. Shyamal "Bird beak adaptations." Available on Wikimedia as BirdBeaksA.svg.)***

a picture of the "biliary tree" in a human liver. This is all good. However, much is missing: What is the pattern of veins and arteries through a human liver? What do different types of liver cancer look like? What does a shark's or a pelican's liver look like? Where is it placed in each animal's abdomen? You could also wish to find out what volcanos on Venus look like, or who are the people portrayed in Rembrandt's *Night Watch*. There are people who know these things; most likely someone lectures every year on comparative anatomy, on Rembrandt's iconography, and on planetary geology—showing slides to their students and pointing at the salient objects—but this knowledge is only minimally expressed in Wikipedia. Unfortunately, experts do not contribute their visual knowledge to Wikipedia. It is easy to see why: posting, annotating, and linking pictures is boring and time-consuming.

In conclusion, the medium of Wikipedia, and more generally of the web, is not working well: while words are well managed on the web, pictures are not. We cannot easily convey visual knowledge, and we cannot use visual queries. Can we change this state of affairs?

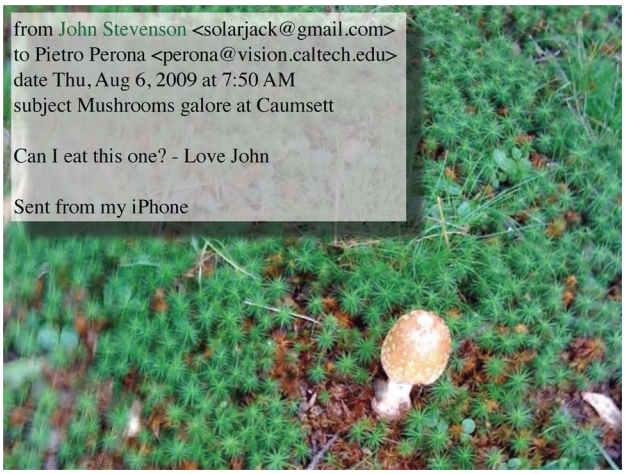


Fig. 3. ***Another mystery picture: which mushroom is this? Is it edible? Try finding out using Wikipedia. Good luck! (Photo courtesy of J. C. Stevenson, who is still alive.)***

III. VISIPEDIA

Like Bush, we could be tempted to dream a bit, and think of an augmented Wikipedia: a "visual encyclopedia" where pictures are first-class citizens alongside text. I will call it *Visipedia* here. Let us start with some examples.

While sitting at the cafe, I could have snapped a picture of one of the pigeons using my camera phone. I could have uploaded the picture to Visipedia as a query. The picture would have become clickable right on my phone, thanks to information sent back by Visipedia. I could have tapped my

finger on the fleshy bit and, *voila!*, obtained the "cere" page of the Wikipedia (see Fig. 4), as well as examples of other birds who have ceres (various hawks and parrots, it turns out).

In another example, a marine ornithologist aboard a research vessel decides to spend the evening organizing her observations of albatrosses. She is happy to share her knowledge, so she edits Visipedia page on "albatross" and uses Visipedia tools to upload and annotate some of the pictures and video she took during the day. Her annotations point out the salient anatomical features, courtship rituals, and how to identify different species of albatross. This is quick work for her since Visipedia tools are highly automated: Visipedia already knows about bird morphology in general and is able to ask the expert intelligent questions by means of graphical overlays; for example, the hook at the end of an albatross' beak and the shape of the nostrils are rather unusual, therefore Visipedia highlights them with a pink shade prompting the expert to name those features, if possible. The expert is relieved from menial work (outlining and annotating in the picture features that are common to many bird species, hyperlinking corresponding structures across species) and can focus on new content. The expert finds this experience productive and enjoyable, and decides to add a page on fulmars as well.

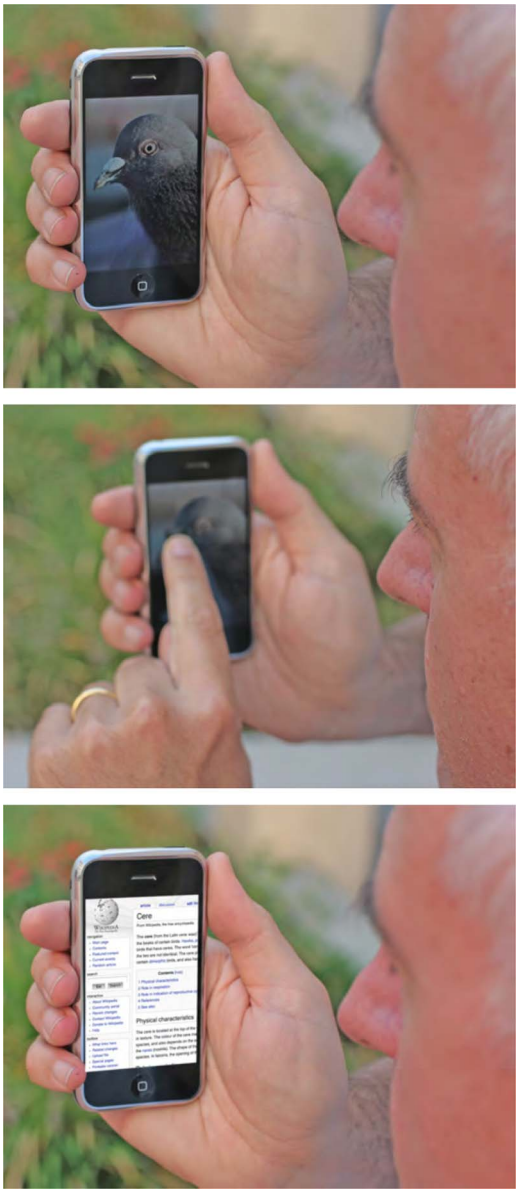


Fig. 4. ***A visual query for Wikipedia. (Top) The user has captured a picture. (Middle) The picture was sent to a central computer which recognizes its contents; as a consequence, meaningful regions of the image become clickable (see also Fig. 6). (Bottom) The relevant Wikipedia page is retrieved.***

As may be seen from these examples, Visipedia generalizes Wikipedia in many ways. Besides hyperlinking text with text and text with pictures, it hyperlinks pictures with pictures and pictures with text. This is technically possible today. We have all encountered web pages where hyperlinks are available from specific "hot" regions of images. Why is this not done in Wikipedia? The answer is that it is too laborious to select all interesting regions by hand on a large corpus of images, and thus it does not happen. Visipedia, thus, will come to life only if a greater degree of automation than Wikipedia is possible. Text is produced word-by-word by people, and thus it is natural (although somewhat boring) to insert hyperlinks by hand and Wikipedia is mostly a handcrafted object (there are already a few *bots* performing menial tasks). In contrast to words, pixels are produced wholesale by machine, and it would be prohibitively time-consuming to insert hyperlinks by hand. Think of footage of an albatross gliding over water: who would like to annotate by hand the anatomical characteristics of its beak in every frame? And that is not all: there are the feet, the wings, the waves, etc. We estimate that there are many tens of meaningful regions even in fairly simple pictures [39]. Thus, in order to make Visipedia a reality, we will need to develop software that can "understand" automatically what is there in pictures, and link related "visual concepts." Humans would, of course, need to be "there" and provide some guidance and information. I will discuss all this in Section V.

Five categories of people will likely interact with Visipedia. As in Wikipedia, there will be *users*, who are interested in finding answers to specific questions and browsing related information via visual hyperlinks; there will be *experts*, who are willing to share what they know with everyone else; there will be also *editors*, nonexperts who will help resolve ambiguities, detect inconsistencies, enforce standards. In addition, there will be *annotators*, "eyeballs for hire" who will help with image segmentation, naming, and other tasks that ultimately will be automated; they will be paid (in cash, kudos, or other currency) to do so (think of Amazon's Mechanical Turk [5fy24z] as an example). There will be one last category

of people: *automation providers*, i.e., computer scientists who will write and upload software for automating tools in Visipedia making it more efficient and useful day-by-day (see Fig. 5).

IV. DESIGN OF VISIPEDIA

What will Visipedia consist of? It could be developed as a "layer" on top of Wikipedia, allowing visual searching. The articles will be the same, and consist of text, pictures, and links. Anyone will be able to add information, e.g., by editing articles, by uploading or annotating a picture. There will be different tiers of users with different degrees of control. But there will be novel aspects as well. The main one is a *high degree of automation*. As I pointed out earlier, experts will not take the time to annotate images in detail unless doing so is easy and quick. Outlining by hand important image regions and hyperlinking them to the appropriate target pages and pictures is rather laborious and boring—this is an excellent expert repellent. Therefore, each image will be first analyzed by machine, component regions will be identified (this process is called "segmentation" by machine vision scientists), objects and other meaningful structures will be identified and measured (answering the "what" and "where" questions), and useful hyperlinks will be assigned automatically (Fig. 6).

Initially, machines will not be able to carry out the complete task of segmentation, understanding, and hyperlinking by themselves. This is for two reasons. First, the definition of many of the regions, as well as their meaning and names, comes from human experience that is not contained in the image itself. Think, for example, of the distinction between "shin" and "calf" in a human leg: the shin is bony and the calf is soft and fleshy. These tactile properties are not evident from photographs of human legs. This distinction must come from a human who has touched and poked legs. However, once an expert provides this information for one image, machines ought to be able to propagate such information to most images of human legs. Second, as I will discuss in Section V, we are not yet able to build automata that can carry out the job.

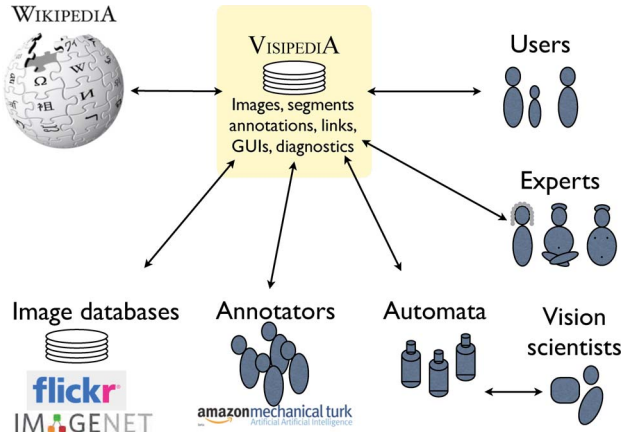


Fig. 5. *A sketch of the structure of Visipedia. Visipedia will provide a link between pictorial queries formulated by a user (top-right) and knowledge on the web, e.g., Wikipedia (top-left) and publicly available image databases (bottom-left). Knowledge will be provided by human experts, as in Wikipedia. Automation will be necessary to analyze images, to link automatically related visual content, to link images and text, and to provide experts with smart user interfaces. This automation will be provided initially by human annotators working behind the scenes (bottom center). Gradually, automata built by machine vision scientists will take over the job (bottom right).*

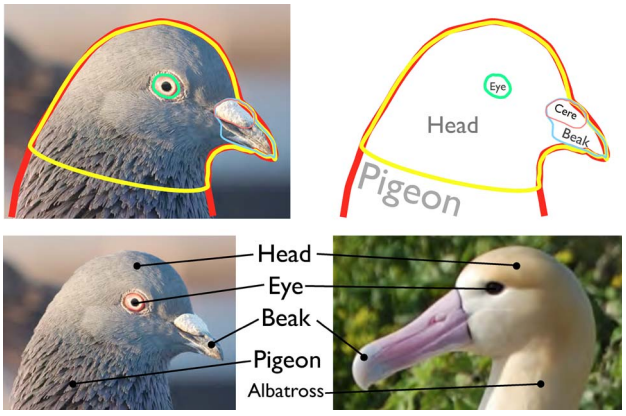


Fig. 6. *(Top) Automated agents in Visipedia will segment the picture of a bird into its component parts, as specified by a human expert. Each part will be named and put in correspondence with corresponding parts of other birds, either recognizing their similarity with previously labeled bird pictures, or with the help of a human annotator. (Bottom) Hyperlinks will be drawn with the corresponding Wikipedia pages, as well as with corresponding parts in other bird pictures. (Albatross picture adapted from Short Tailed Albatross by James Lloyd, available on Wikipedia as Short_tailed_Albatross1.jpg.)*

To experts, Visipedia will initially look like an intelligent graphical user interface for annotating images, helping them share their visual knowledge with other humans. In time, Visipedia may also take the form of an intelligent pupil, asking lots of questions and making the teachers aware of inconsistencies and gaps in the knowledge they provided. Much of our visual knowledge is, after all, implicit; a smart pupil asking lots of questions is the best way to extract that hidden knowledge.

While experts will provide and annotate some paradigmatic images, there is also much implicit knowledge available on the web, which could help Visipedia go beyond the contribution of an individual. Visipedia's automata will access the many large public repositories of images and attempt to annotate those images automatically based on the expert-provided templates. For example, as soon as an expert has annotated the picture of one pelican, all good pelican images on the web (e.g., those uploaded by birders to Flickr) should be annotated automatically; their anatomical landmarks should be put in correspondence with those of other birds and animals; the template

provided by the pelican expert should be compared with that provided by the cormorant expert. The experts will be asked for help resolving difficult cases, e.g., when new structure is revealed by looking at the bird from a different viewpoint, and when templates provided by different experts do not correspond. The expert will thus be tapped to provide further useful information, and Visipedia will be able to generalize from the original example, and learn about the natural variability in a given domain.

Another essential aspect of Visipedia is that the processes that analyze images automatically, and link them to text and other images, will self-diagnose. They must identify unknown and ambiguous bits that need to be resolved by either a human expert or by another automated agent. Not all such questions are equally urgent and useful. Visipedia processes must be able to estimate which questions are most informative and submit those to human experts, a bit like playing 20 questions [4], [22], [25]. This way, human experts will be asked interesting, rather than boring, questions, and will feel that their time is well spent. Some may even enjoy the process, a bit like playing a game.

Alongside deciding which questions to ask, Visipedia will have to decide whom to ask. It will have to organize human annotators, experts, and automata according to subject of expertise, and according to the depth of that expertise. It will have to have means of crosschecking answers, perhaps by posing the same question to more than one agent. For example, Visipedia will estimate the reliability of humans, difficult as that may be, as well as that of automata. From the point of view of Visipedia, there is little difference of which is which.

Which work will be done by humans and which by software? One could conceive of an early incarnation of Visipedia where much of the busy work is done by humans, behind the scenes. These humans will either have to be paid, e.g., via Amazon's Mechanical Turk [38], [39], [43], [44], or will do the work in exchange for some benefit, as in LabelMe [35], or for fun, as in Murray's first *Oxford English Dictionary* [49], and in the recent ESP game [46]. Relying on human labor will initially limit the scope and size of Visipedia; in time, the boundary between manual labor and automation will recede, and more and more work will be taken over by automatic processes. Thus, Visipedia will accommodate multiple agents, as does today's Wikipedia, with the difference that some of the agents will be automatic. Like today, the absolute and relative reliability of different agents will have to be assessed, disagreements arbitrated, and conflicts resolved; this will be done by a combination of people and automated agents.

Who will provide the automation on which Visipedia will be built? Who will write software for segmenting images, for recognizing objects and categories, for deciding which patches of which image link to a given word or image region? Automating image analysis is very much of a research topic, as I will discuss in Section V. It would be highly unlikely that a single person, or a group of people, would be sufficiently clever and wise to produce the necessary software all at once, or even over an extended period of time. It will be far more productive if, like articles in Wikipedia, contributing automated agents to Visipedia is open to all. Anyone in the world should be able to contribute automation software to Visipedia, and such software should live or die only by virtue of its usefulness. This points, once again, to the need for processes that measure the quality of human and automatic agents alike, and combine information from these agents according to their reliability.

V. MACHINE VISION FOR VISIPEDIA

Automation is perhaps the most interesting technical aspect of Visipedia. Human–computer interfaces, automated natural language understanding, machine learning, knowledge representation, and automated reasoning will all play an important role. The workhorse of this automation is *machine vision*, whose purpose is to understand the content of images. What does "understanding an image" mean? Marr [31] defined "vision" as *"knowing what is where by looking."* Indeed, we see many "things," shapes, spatial and causal relationships, and actions in a picture, even in a brief glance [13]. The goal of machine vision is precisely this: to produce a synthetic description of what is meaningful and relevant in each picture [39], and to relate it to other pictures and text [1]. To give a few examples, it would be useful if Visipedia could recognize automatically the main objects that are present in an image, either as individuals (e.g., Michelangelo's *David*) or as members of categories (e.g., a frog), the nature and the shape of surfaces (e.g., "marble sphere"), the identity and category of scenes (e.g., "renaissance square"), the weather and time of the day (e.g., "hazy morning"), and the authorship and subject of paintings (e.g., "portrait of a man by Bellini"). In video one would like to recognize actions and activities carried out by humans and animals, trajectories of objects, interactions (e.g., "an angry woman scolding a child"). Clearly, reading text will also be useful. Machine vision automata working for Visipedia will integrate the knowledge of many human experts, and therefore, will surpass the visual competence of any single human in interpreting an image.

Automating image understanding is an achievable goal. Machine vision scientists have been making progress by leaps and bounds during the relatively short life of this field. Much is understood now on how to segment an image into component regions [33], [34], how to compute the 3-D layout of a scene from images and video [2], [21], [28], [36], how to recognize individual objects [29] and object categories [47], and how to detect faces [45], people [7], surfaces [41], scenes [14], and actions and activities [8], [23]. Researchers are studying approaches to combining these

cues and processes into a single "image understanding" system [20], [27], [40]. Large-scale efforts are under way to collect a body of visual knowledge that will allow our algorithms to learn what goats, fire extinguishers, and ice cream cones look like, and where they typically appear in pictures [9], [17], [35], which words are typically associated with which pictures and objects [1], [11], [19], [26], and how to discover images that look alike, in order to link them together [24], [30]. I am just giving here a few references, for the sake of readers being able to get a sample of the literature, but the corpus of work on these topics is quite enormous, and each topic has many interesting facets. Indeed, a great number of talented people, most of them young, are working at universities and corporate laboratories around the world to solve the problem of automating different aspects of vision. Readers will get a good feel for what is possible by leafing through the pages of this issue of the Proceedings of the IEEE.

Is the state of the art in machine vision sufficient to realize Visipedia? The short answer is no (but I have a "yes" answer as well; you will find it in the next paragraph). The answer is "no" for two reasons. First, the performance of our algorithms is not yet good enough for most applications, e.g., our best algorithms for detecting humans in pictures can detect only a fraction of all pedestrians in pictures of urban scenes [10]. There are many tasks for which we do not yet have performance figures and it is tempting (and wise) to assume that there is even a greater number of useful tasks that we have not even begun thinking about. Second, we have not yet tried to build systems, such as Visipedia, where a number of heterogeneous machine vision algorithms have to join forces to interpret an image fully. There are so many unexplored failure modes in such a complex system! So, no, we are nowhere near a fully automated image understanding system to power Visipedia, and we are not even sure if we have a fair understanding of the difficulties that lie ahead of us.

And yet, the answer could also be "yes." Building Visipedia should be approached incrementally, by adding useful automated agents, one by one, to Wikipedia. One does not need to solve the problem all at once. For example, detecting human faces is now not only a fairly well-understood academic problem [45], but it has made its debut in industrial products (Picasa in 2008 [5fvpxz], and Apple iPhoto in 2009 [cy9no4]). It does not work perfectly, but it is good enough for some applications; furthermore, a module does not need to be perfect to be useful: even if not all faces were detected automatically, it would be useful to have detected a majority of them and let humans complete the job where this is needed. Some steps are best carried out by combining automation with a little human guidance [34]. Surely, detecting faces in pictures will be a useful addition to Wikipedia; of course, someone needs to build a convenient graphical user interface to allow Wikipedia editors and users to add name tags to the pictures. This first step could be realized today. Similarly, useful incremental steps will follow as we perfect human detectors, scene classifiers, image segmentation, etc. There are functioning and publicly available machine vision modules (see, e.g., the Open Computer Vision Library [5] and VLFeat [42]) that one could stitch together into a working first implementation of Visipedia. How far would that go? We will not know until we try.

Is this, then, the right time to try and build Visipedia? Yes, provided that we set realistic goals for the first incarnation of Visipedia. The first system should not aim to be useful for the users of Wikipedia at large. One should pick a well-defined domain, e.g., "birds" or "plants" [3], [32], with a community of highly motivated enthusiasts, and provide some set of tools and functionalities that will allow users to start accumulating visual knowledge online. This initial seed effort will suggest ways in which Visipedia can grow, and will point machine vision researchers (and their collaborators in machine learning, human–machine interaction, knowledge management, etc.) to the most productive fundamental research problems. Thinking about Visipedia is, in itself, an attempt to narrow the general problem of "pictures on the web" to a manageable subproblem that one can think and do something about.

Are there any aspects that would be crucial to putting together Visipedia and have been, so far, ignored by machine vision researchers? Self-diagnosing, active incremental learning, and human–machine interaction come to mind. As we have seen in Sections III and IV, Visipedia will be the result of the cooperation of human and automated agents. A key requirement for inducing humans to contribute their knowledge is making their work expeditious and, hopefully, fun. To avoid boring humans and wasting their time, automated agents will need to decide which bits of knowledge so far collected are most "inconclusive" or "ambiguous," so that only the most informative questions would be asked [22], [25], [43], [44], [48].

Unlike the static benchmark data sets that vision researchers are used to, the web is a dynamic environment: automata will not be able to wait until all the data are in before learning something useful. When new data show up, they must be digested quickly without having to reconsider the old training data. Thus, learning must happen incrementally [12], [16], very much as it happens in animals and humans. Another underexplored challenge is scaling to large data sets, to billions of training examples, and to hundreds of thousands of categories [15], [18], [37]. Knowledge will have to be incorporated incrementally [16]. Much needs to be done before automated agents can interact with humans in real time.

How will Visipedia come about? Who will take the step of making it happen? Will it be the result of a loosely coordinated effort of researchers, users, and experts and, like Wikipedia, open to all and entirely free? Will it come out of Google, Microsoft, or some other large corporation? My guess is that Visipedia is more likely to result from a colony of heterogeneous cooperating agents, rather than a

monolithic design; it is more likely to succeed if hundreds of researchers around the world are empowered to upload their latest and best software. Therefore, I am betting on an open architecture winning the race. A central kernel will have to be designed and implemented by a small and clever team—both a large corporation and an academic group could contribute this initial seed.

VI. CONCLUSION

The web, as wonderful as it may be, is not perfect. Some queries and some knowledge are best expressed with pictures, not in words. Unfortunately, searching, indexing, organizing, and hyperlinking pictures (photographs, drawings, video) by their content, rather than by the text that surrounds them, is not yet possible. As a result, pictures, the vast majority of the data we store, are a sort of digital dark matter: they are there but they are not accessible. Conversely, human experts carry with them much visual knowledge that they are not able to share on the web with other humans.

Visipedia, making pictures (photographs, video, drawings) on Wikipedia as searchable as text and empowering people to contribute their visual expertise, will make it easier to collect and share human knowledge. Visipedia appears to be a realistic goal for the next ten years. Automating vision and finding ways to allow large numbers of humans and automated agents to interact successfully appear to be the main challenges.

Visipedia is just an example of what will be possible when functioning machine vision algorithms become widely available and will be integrated with humans, pictures, and text on the web. The ability to annotate, search, and organize visual and verbal knowledge automatically will change many aspects of science, medicine, manufacturing, security, education, information, and entertainment.

Machine vision researchers can play a major role in transforming the web and making Visipedia a reality. Besides attempting to solve the fundamental technical challenges of automating image understanding, we must engage in building end-to-end systems where machine vision and other automated agents collaborate seamlessly with human users, experts, annotators, and editors. The system-building effort should proceed in parallel with the more fundamental research: without a set of concrete goals it will be difficult to aim the basic research in the right direction. ■

Acknowledgment

This paper was written while the author was visiting the Department of Information Engineering (DEI), University of Padova, Padova, Italy. He would like to thank Profs. G. Picci and G. Cortelazzo for their warm hospitality. The concept of Visipedia was developed in collaboration with T. Mita and P. Welinder at Caltech, and with S. Belongie and his students at the University of California San Diego (UCSD). The author would like to thank many colleagues who helped him formulate and present the ideas expressed in this paper. C. Tomasi made the author aware of Bush's memex. J. Stevenson, K. Grauman, K. Branson, and P. Dollar read a draft and improved both language and exposition. L. Lazebnik, A. Efros, and F.-F. Li made many useful comments. The concept was presented in August 2009 at Banff, AB, Canada, during the "Vision and the Internet" workshop. A number of participants made useful suggestions and comments, in particular, J. Malik, K. Koutulakos, D. Forsyth, B. Aguera y Arcas, R. Szeliski, D. Hoiem, S. Seitz, and A. Zisserman.

REFERENCES

[1] K. Barnard, P. Duygulu, D. A. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

[2] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vis.*, vol. 12, no. 1, pp. 43–77, 1994.

[3] P. N. Belhumeur, D. Chen, S. Feiner, D. W. Jacobs, W. J. Kress, H. Ling, I. Lopez, R. Ramamoorthi, S. Sheorey, S. White, and L. Zhang, "Searching the world's herbaria: A system for visual identification of plant species," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 116–129.

[4] G. Blanchard and D. Geman, "Sequential testing designs for pattern recognition," *Ann. Stat.*, vol. 33, no. 3, pp. 1155–1202, 2005.

[5] G. R. Bradski, *Learning OpenCV: Computer Vision With the OpenCV Library*. Sebastopol, CA: O'Reilly, 2008.

[6] V. Bush, "As we may think," *The Atlantic Monthly*, vol. 176, pp. 101–108, Jul. 1945.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[8] H. Dankert, L. Wang, E. D. Hoopfer, D. J. Anderson, and P. Perona, "Automated monitoring and analysis of social behavior in drosophila," *Nature Methods*, vol. 6, no. 4, pp. 297–303, Apr. 2009.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[10] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 304–311.

[11] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1778–1785.

[12] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2005.

[13] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?" *J. Vis.*, vol. 7, no. 1534–7362 (electronic), pp. 1–29, 2007.

[14] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 524–531.

[15] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *Proc. Int. Conf. Comput. Vis.*, 2005, pp. 1816–1823.

[16] R. Gomes, M. Welling, and P. Perona, "Incremental learning of nonparametric Bayesian mixture models," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2008, DOI: 10.1109/CVPR.2008.4587370.

[17] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Pasadena, Tech. Rep. CNS-TR-2007-001, 2007.

[18] G. Griffin and P. Perona, "Learning and using taxonomies for fast visual categorization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2008, DOI: 10.1109/CVPR.2008.4587410.

[19] S. Gupta, J. Kim, K. Grauman, and R. J. Mooney, "Watch, listen & learn: Co-training on captioned images and videos,"

in *Proc. Eur. Conf. Mach. Learn. Knowl. Disc. Databases*, 2008, pp. 457–472.

[20] F. Han and S. C. Zhu, "Bottom-up/top-down image parsing with attribute graph grammar," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 59–73, Jan. 2009.

[21] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *Int. J. Comput. Vis.*, vol. 80, no. 1, pp. 3–15, 2008.

[22] A. D. Holub, M. C. Burl, and P. Perona, "Entropy-based active learning for object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit./2nd IEEE Workshop Online Learn. Classification*, 2008, DOI: 10.1109/CVPRW.2008.4563068.

[23] N. Ikizler and D. A. Forsyth, "Searching for complex human activities with no visual examples," *Int. J. Comput. Vis.*, vol. 80, no. 3, pp. 337–357, 2008.

[24] P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman, "Online metric learning and fast similarity search," in *Proc. Neural Inf. Process. Syst.*, 2008, pp. 761–768.

[25] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Active learning with Gaussian processes for object categorization," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2007, DOI: 10.1109/ICCV.2007.4408844.

[26] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 951–958.

[27] L.-J. Li, G. Wang, and L. Fei-Fei, "Optimol: Automatic online picture collection via incremental model learning," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2007, DOI: 10.1109/CVPR.2007.383048.

[28] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, pp. 133–135, 1981.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[30] T. Malisiewicz and A. A. Efros, "Beyond Categories: The visual memex model for reasoning about object relationships," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Dec. 2009.

[31] D. Marr, *Vision*. San Francisco, CA: Freeman, 1982.

[32] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. Indian Conf. Comput. Vis. Graph. Image Process.*, 2008, pp. 722–729.

[33] X. Ren, C. Fowlkes, and J. Malik, "Learning probabilistic models for contour completion in natural images," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 47–63, 2008.

[34] C. Rother, V. Kolmogorov, and A. Blake, "'Grabcut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[35] B. C. Russell, A. B. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 157–173, 2008.

[36] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, 2002.

[37] J. Sivic and A. Zisserman, "Video google: Efficient visual search of videos," in *Toward Category-Level Object Recognition*, vol. 4170, J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, Eds. Berlin, Germany: Springer-Verlag, 2006, pp. 127–144.

[38] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis Pattern Recognit.*, 2008, DOI: 10.1109/CVPRW.2008.4562953.

[39] M. Spain and P. Perona, "Some objects are more equal than others: Measuring and predicting importance," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 523–536.

[40] Z. W. Tu, X. R. Chen, A. L. Yuille, and S. C. Zhu, "Image parsing: Unifying segmentation, detection and recognition," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 113–140, 2005.

[41] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vis.*, vol. 62, no. 1–2, pp. 61–81, 2005.

[42] A. Vedaldi, *Vlfeat, an Open-Source Computer Vision Library*. [Online]. Available: http://www.vlfeat.org/

[43] S. Vijayanarasimhan and K. Grauman, "Multi-level active prediction of useful image annotations for recognition," *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, Dec. 2008.

[44] S. Vijayanarasimhan and K. Grauman, "What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2262–2269.

[45] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

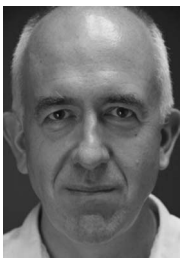
[46] L. von Ahn, "Games with a purpose," *IEEE Comput. Mag.*, vol. 39, no. 6, pp. 96–98, Jun. 2006.

[47] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *Proc. Eur. Conf. Comput. Vis.*, Dublin, Ireland, 2000, pp. 18–32.

[48] P. Welinder and P. Perona, "Active discrimination: An online algorithm for finding experts and obtaining cost-effective labels," in *Proc. Workshop Adv. Comput. Vis. With Humans in the Loop at Comput. Vis. Pattern Recognit.*, pp. 1–8, 2010.

[49] S. Winchester, *The Professor and the Madman: A Tale of Murder, Insanity, and the Making of The Oxford English Dictionary*. New York: Harper Collins, 1998.

ABOUT THE AUTHOR



Pietro Perona is a graduate of the University of Padova, Padova, Italy, in 1985 and received the Ph.D. degree in electrical engineering and computer science from the University of California at Berkeley, Berkeley, in 1990.

Before joining California Institute of Technology (Caltech), Pasadena, in 1991, he was a Postdoctoral Fellow with the International Computer Science Institute (ICSI), University of California at Berkeley and with the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology (MIT), Cambridge. Currently, he is the Allen E. Puckett Professor of Electrical Engineering at Caltech. He is also the Executive Officer of the Computation and Neural Systems Ph.D. program at Caltech. He is interested in computational vision, machine vision, and human visual perception. He has recently become interested in modeling and measuring animal behavior. His work includes partial differential equations for early visual processing (*anisotropic diffusion*), visual psychophysics and modeling of texture boundary perception, shape from shading, attention and recognition, as well as computational models of visual categorization and learning (the *constellation model*).